Canalizing structure of genetic network dynamics: modelling and identification via mixed-integer programming

Cinquemani, Eugenio ; Porreca, Riccardo ; Lygeros, John ; Ferrari-Trecate, Giancarlo

Abstract: We discuss the identification of genetic networks based on a class of boolean gene activation rules known as hierarchically canalizing functions. We introduce a class of kinetic models for the concentration of the proteins in the network built on a family of canalizing functions that has been shown to capture the vast majority of the known interaction networks. The simultaneous identification of the structure and of the parameters of the model from experimental data is addressed based on a mixed integer parametrization of the model class. The resulting regression problem is solved numerically via standard branch-and-bound techniques. The performance of the method is tested on simulated data generated by a simple model of Escherichia coli nutrient stress response.

DOI: https://doi.org/10.1109/CDC.2009.5400670

Canalizing structure of genetic network dynamics: modelling and identification via mixed-integer programming

Eugenio Cinquemani, Riccardo Porreca, John Lygeros and Giancarlo Ferrari-Trecate

***Abstract*— We discuss the identification of genetic networks based on a class of boolean gene activation rules known as hierarchically canalizing functions. We introduce a class of kinetic models for the concentration of the proteins in the network built on a family of canalizing functions that has been shown to capture the vast majority of the known interaction networks. The simultaneous identification of the structure and of the parameters of the model from experimental data is addressed based on a mixed integer parametrization of the model class. The resulting regression problem is solved numerically via standard branch-and-bound techniques. The performance of the method is tested on simulated data generated by a simple model of *Escherichia coli* nutrient stress response.**

I. INTRODUCTION

Genetic network modelling and identification have been addressed at different levels of detail, depending on the available data. Qualitative interaction models have been considered in [8] in the form of Boolean networks, and in [14], [15] in the form of Bayesian networks. Such discrete approaches are well suited for systems observed at equilibria, but many interactions and the causality of the interactions may be obscured.

An alternative approach is to use kinetic models based on ODEs, see e.g. [9]. In principle, a kinetic model can be fitted to time-course data, yielding a complete view of the regulation network. However, an overwhelming number of models need to be parsed. A partial remedy is to turn the problem into black-box parametric identification, i.e. via neural networks [22].

A middle ground between discrete and kinetic models is provided by linearization methods [17], [18], [19]. A linearized version of a kinetic model is fitted to several perturbed equilibria of the system. This provides hints on the presence and the strength of the interactions among genes, but the assumption that the linearized dynamics are the same at all equilibria makes these methods somewhat limited. A stochastic approach that exploits intrinsic gene expression noise as an inherent perturbation signal has been proposed in [23].

A promising alternative is hybrid modelling [20], [21], [1]. Full-blown kinetic models are turned into piecewise linear models by the use of step activation functions. Thus, different linear models are fitted to the data over different partitions of the state space. Unfortunately, the approximation of activation functions via step functions is too coarse for certain systems. A stochastic hybrid approach was explored in [24], where activation functions have been replaced by sigmoidal switching probabilities for the genetic regulatory events.

The objective of this work is to develop an efficient approach to genetic network identification via kinetic modelling. In the context of Boolean networks, it was observed in the literature that most of the known gene activation rules fall in the class of Hierarchically Canalizing Functions (HCF) [5], [6], [7]. Our aim is to translate this knowledge into the context of kinetic modelling and exploit it for the structural and parametric identification of genetic network dynamics. Ideas on how to apply HCF in the context of Boolean networks are discussed e.g. in [11], [12], [13].

Boolean gene activation rules and HCF are reviewed in Section II. Kinetic modelling of gene activation dynamics is discussed in Section III. In Section IV we state the genetic network identification problem in a class of ODE models with HCF-like structure. A formulation of the identification problem in terms of mixed-integer optimization is developed in Section V. Performance is assessed in Section VI by numerical experiments on a simulated model of nutrients stress response in bacterium *Escherichia coli*.

II. BOOLEAN RULES: HIERARCHICALLY CANALIZING FUNCTIONS

Boolean network modelling [8] describes the activation of each gene as a boolean function of the expression of the genes in network. Consider a network with n genes. For $i = 1, \dots n$, let $X_i \in \{0, 1\}$ indicate the activation status of gene i at a fixed time instant; $X_i = 1$ means that gene i is expressed, while $X_i = 0$ means it is not. The activation of gene i at the next time instant is indicated by X_i^+ and is modelled by

$$X_i^+ = b_i(X_1, \dots, X_n),$$

where the function $b_i : \{0, 1\}^n \to \{0, 1\}$ is a *boolean rule*.

Canalizing functions are a subclass of boolean rules whose output is determined by at least one value of at least one input variable. That is, there exists a *canalizing input* X_j such that, if X_j takes the *canalizing value* $U \in \{0, 1\}$, then the function $b_i(X_1, \dots, X_n)$ takes on the *canalized value* $Z \in \{0, 1\}$ regardless of the value of the other variables X_k, with $k \neq j$. Within this class, Hierarchically Canalizing Functions (HCFs) with $\ell \leq n$ effective inputs are characterized by the existence of an ordered subset $(X_{j_1}, \dots, X_{j_\ell})$ of the

This work was supported in part by the SystemsX.ch research consortium under the project YeastX.

Eugenio Cinquemani and John Lygeros are with the Institut für Automatik, ETH Zürich, Switzerland.

Riccardo Porreca and Giancarlo Ferrari-Trecare are with the Dipartimento di Informatica e Sistemistica, Università di Pavia, Italy.

Corresponding author: Eugenio Cinquemani, email: `cinquemani@control.ee.ethz.ch`

variables $X_1, \ldots, X_n$ with the following property. Input X_{j_1} is canalizing with canalizing value U_{j_1}. When X_{j_1} takes its non-canalizing value $1-U_{j_1}$ (i.e. its Boolean negation), X_{j_2} is canalizing with canalizing value U_{j_2}. When both X_{j_1} and X_{j_2} take on their non-canalizing values, X_{j_3} is canalizing, an so on. Finally, when $X_{j_1}, X_{j_2}, \ldots, X_{j_{\ell-1}}$ take on their non-canalizing values, the value of the expression is determined by X_{j_ℓ}.

HCFs received much attention since it was observed that the vast majority of the known regulatory interactions among genes can be written in terms of HFC-type rules, see [6], [5], [7]. In particular, two specific subfamilies of HCFs, termed S_0^ℓ and S_1^ℓ, appear to explain most of the known interactions. For a given variable X_j, let $\sigma^\pm(X_j)$ be either X_j or $1-X_j$. Then, for some $\ell \leq n$ and pairwise different indices $j_1, j_2, \ldots, j_\ell$ from the set $\{1, \ldots, n\}$, $b_i(X_1, \ldots, X_n)$ is equal to

$$\sigma^\pm(X_{j_1}) \cdot \ldots \cdot \sigma^\pm(X_{j_\ell}), \tag{1}$$

when $b_i \in S_0^\ell$, and to

$$\sigma^\pm(X_{j_1}) \cdot \ldots \cdot \sigma^\pm(X_{j_{\ell-2}}) \cdot \big(1 - \sigma^\pm(X_{j_{\ell-1}})\sigma^\pm(X_{j_\ell})\big), \tag{2}$$

if $b_i \in S_1^\ell$ (see [2]). Note that functions in S_0^ℓ correspond to chains of "and" operations, whereas functions in S_1^ℓ also include an "or" (rightmost factor of (2)).

III. KINETIC MODELS OF GENETIC NETWORKS

For $t \in \mathbb{R}$ and $i = 1, \ldots, n$, let $x_i(t) \in \mathbb{R}_+$ denote the concentration at time t of the protein encoded by gene i. Consider the following dynamical model:

$$\dot{x}_i = -\gamma_i x_i + g_i(x_1, \ldots, x_n), \tag{3}$$

where $\gamma_i \in \mathbb{R}_+$ is the unregulated degradation constant. Function $g_i : \mathbb{R}_+^n \to \mathbb{R}$ represents a variable synthesis rate that encodes the regulatory effects of all proteins in the network on the expression of gene i. In practice, only a subset of these proteins act as transcription factors for gene i. In this case, g_i effectively depends on a subset of $\{x_1, \ldots, x_n\}$.

Kinetic-type models (see e.g. [9]) express $g_i(x_1, \ldots, x_n)$ as a weighted combination (sums and products) of sigmoidal functions $\sigma^+(x_j, \theta)$ or $\sigma^-(x_j, \theta)$, with $j = 1, \ldots, n$, where $\sigma^+(\cdot, \theta) : \mathbb{R}_+ \to [0, 1]$ is increasing, parameter vector θ determines the "shape" (e.g. threshold and steepness) of the sigmoid, and $\sigma^- = 1 - \sigma^+$. In this case, $g_i(x_1, \ldots, x_n)$ represents the activation level of the expression of the gene.

Example 1: The response of the bacterium *Escherichia coli* to changes in the availability of carbon was modelled in detail in [4]. A starvation (input) signal $u_s : \mathbb{R} \to \mathbb{R}_+$ indicates abundance ($u_s = 0$) or lack ($u_s = 1$) of carbon, and determines the activation or deactivation of the response-to-starvation mechanism. For the case of carbon abundance ($u_s = 0$), the model is:

$$\dot{x}_1 = \kappa_1^1 + \kappa_1^2 \ - \gamma_1 x_1, \tag{4}$$

$$\dot{x}_2 = \kappa_2^1 + \kappa_2^3 \sigma^-(x_3, \theta_3^1) - \gamma_2 x_2, \tag{5}$$

$$\dot{x}_3 = \kappa_3^1 \sigma^-(x_3, \theta_3^5) + \kappa_3^2 \ \sigma^+(x_4, \theta_4^1)\, \sigma^-(x_5, \theta_5^2)\, \sigma^-(x_3, \theta_3^5) - \gamma_3 \ x_3, \tag{6}$$

$$\dot{x}_4 = \kappa_4 (1 - \sigma^+(x_4, \theta_4^2)\, \sigma^-(x_5, \theta_5^1)) \ \sigma^-(x_3, \theta_3^4) - \gamma_4 x_4, \tag{7}$$

$$\dot{x}_5 = \kappa_5 \ \sigma^+(x_4, \theta_4^2)\, \sigma^-(x_5, \theta_5^1)\, \sigma^+(x_3, \theta_3^4) - \gamma_5 \ x_5, \tag{8}$$

$$\dot{x}_6 = \kappa_6^1 + \kappa_6^2 \sigma^+(x_3, \theta_3^3) - \gamma_6 x_6. \tag{9}$$

Subscripts 1 through 6 stand for proteins Cya, CRP, Fis, GyrAB, TopA and rrn, in the same order. Each θ_i is a vector of parameters for a sigmoid acting on the i-th protein concentration. Different superscripts indicate possibly different parameter values. Coefficients $\kappa_i \in \mathbb{R}_+$ quantify the protein synthesis rate when gene i is expressed. Different superscripts correspond to alternative activation paths and/or baseline production rates. A detailed illustration of the full model can be found in [1].

Note that the algebraic structure of all activation functions in Example 1 is similar to that of (1)–(2). We shall formalize this in the next section to include information about the canalizing structure of gene interaction networks in a kinetic modelling framework. The only exception is the equation for x_3, which shows the sum of two terms. We will come back to this equation in Section VI.

IV. PROBLEM STATEMENT

Given a regulatory network with n genes, we address the problem of identifying a model for the dynamics of the system in the form (3). We assume that noisy measurements $y_1(t), \ldots y_n(t)$ of the concentrations $x_1(t), \ldots, x_n(t)$ are available for $t \in \mathscr{T}$, where $\mathscr{T} = \{t_1, t_2, \ldots, t_m\}$ is a sequence of observation instants. We further assume that noisy measurements $s_1(t), \ldots, s_n(t)$ of the synthesis rates $g_1, \ldots, g_n$ are available at the same time instants $t \in \mathscr{T}$. It is shown e.g. in [16] that time-course concentration measurements and the relevant synthesis rates can be drawn from dedicated experiments under appropriate experimental conditions. For the observations we shall consider a model with multiplicative noise,

$$y_i(t) = \big(1 + e_i(t)\big) x_i(t), \tag{10}$$

$$s_i(t) = \big(1 + \epsilon_i(t)\big) g_i(t), \tag{11}$$

where $\{e_i(t) : \ t \in \mathscr{T}, \ i = 1, \ldots, n\}$ and $\{\epsilon_i(t) : \ t \in \mathscr{T}, \ i = 1, \ldots, n\}$ are i.i.d random variables with mean 0 and variance $\sigma_e^2 > 0$ and $\sigma_\epsilon^2 > 0$. This model appears to be well suited to describe nonnegative observations of protein concentrations [25].

For $i = 1, \ldots, n$, we wish to solve the following regression problem:

$$\min J, \quad J = \sum_{h=1}^{m} w_h \Big(s_i(t_h) - g_i\big(y_1(t_h), \ldots, y_n(t_h)\big) \Big)^2, \tag{12}$$

where the weights $w_h \in \mathbb{R}_+$ must be chosen based on the relative accuracy of the estimates, i.e. on the noise model, and the minimum is taken with respect to a suitable class of gene activation functions. In light of the discussion of Section II, we make the following assumption.

Assumption 1: For $i = 1, \dots, n$, there exists a nonnegative integer $\ell \leq n$ such that

$$g_i(x_1, \dots, x_n) = \kappa_i^1 + \kappa_i^2 b_i(x_{j_1}, \dots, x_{j_\ell}),$$

where $\kappa_i^1 \in \mathbb{R}_+$, $\kappa_i^2 \in \mathbb{R}_+$, $(j_1, \dots, j_\ell)$ is an ordered subset of $\{1, \dots, n\}$ and $b_i(x_{j_1}, \dots, x_{j_\ell})$ takes one of the two forms

$$\prod_{k=1}^{\ell} \sigma^{\pm}(x_{j_k}, \theta_{i,j_k}),$$

$$\prod_{k=1}^{\ell-2} \sigma^{\pm}(x_{j_k}, \theta_{i,j_k}) \cdot \big(1 - \sigma^{\pm}(x_{j_{\ell-1}}, \theta_{i,j_{\ell-1}}) \sigma^{\pm}(x_{j_\ell}, \theta_{i,j_\ell})\big).$$

In turn, each $\sigma^{\pm}(x_j, \theta_j)$ is either a positive (σ^+) or a negative (σ^+) sigmoid with parameters θ_j.

With this assumption we introduce the HCF structure given by (1)–(2) into kinetic modelling. With abuse of terminology, we will still speak about S_0^ℓ and S_1^ℓ models to mean kinetic models in one of the two forms above. (For $\ell = 0$, we assume that $S_0^0 = S_1^0 = \{b_i\}$ with $b_i \equiv 1$.) The assumption allows us to largely reduce the complexity of the identification problem. This amounts to minimizing J with respect to the functions $b_i \in S_0^\ell \cup S_1^\ell$, with $0 \leq \ell \leq n$, and all the unknown parameters $\theta_{i,j}$. Identification may result in a poor matching of the data in the few outstanding cases where the gene activation rule is not in $S_0^\ell \cup S_1^\ell$. This situation will not be addressed here.

A. Choice of the regression weights

Weights attributed to the error terms in J must reflect the reliability of the data. In principle, weights should account for the uncertainty from s_i and from $g_i\big(y_1(t_h), \dots, y_n(t_h)\big)$. The noise contribution from $y_1(t_h), \dots, y_n(t_h)$ is reshaped by the unknown function g_i and is difficult to quantify. For the time being we shall ignore this contribution and discuss this approximation in light of numerical results in Section VI. The standard deviation of the $\epsilon_i(t_h)$ is proportional to $g_i(t_h)$. For σ_ϵ not too large, the latter will be of the same order as $s_i(t_h)$. Therefore we define the weights as $w_h \propto s_i^{-2}(t_h)$.

B. Model complexity and overfitting

It is well known that matching data based on a class of arbitrarily complex models may result in overfitting. To counteract this we modify the cost function by including a term that penalizes overly complicated models. Several criteria exist and are accompanied by solid theoretical guarantees. For numerical convenience, we opt for the Final Prediction Error (FPE) principle [10], which results in the modified optimization problem

$$\min\ J', \quad J' = \begin{cases} \frac{m+p}{m-p} \times J, & \text{if } p < m, \\ +\infty, & \text{otherwise,} \end{cases} \tag{13}$$

where p is the number of the optimization parameters (unknown parameters of the model). This number depends on the parametrization of the model class and on the specific form of the sigmoidal activation functions. Recall that m is the number of data points.

V. FORMULATION AS A MIXED INTEGER OPTIMIZATION PROBLEM

Direct solution of the regression requires an exhaustive exploration of the class of rules $S_0^\ell \cup S_1^\ell$ and is too costly. To ameliorate this limitation, we recast the problem in a mixed-integer optimization framework. This allows one to solve the problem by well-established branch-and-bound techniques. While the complexity of the worst-case solution remains unchanged, existing branch-and-bound heuristics tend to result in a very effective exploration of $S_0^\ell \cup S_1^\ell$, $\ell = 0, \dots, n$, with enormous savings in terms of computational cost.

Let $\alpha = (\alpha_1, \dots, \alpha_n)$, $\alpha^* = (\alpha_1^*, \dots, \alpha_n^*)$ and $\beta = (\beta_1, \dots, \beta_n)$ be in $\{0, 1\}^n$. Define $\tilde{g}_i = \kappa_i^1 + \kappa_i^2 \tilde{b}_i$ with

$$\tilde{b}_i = \left[\prod_{j=1}^{n} \left[(1 - \alpha_j) + \alpha_j \Big(\beta_j \sigma_{i,j}^+ + (1 - \beta_j)\sigma_{i,j}^-\Big)\right]\right] \times$$

$$\left[1 - \prod_{j=1}^{n} \left[(1 - \alpha_j^*) + \alpha_j^* \Big(\beta_j (1 - \sigma_{i,j}^+) + (1 - \beta_j)(1 - \sigma_{i,j}^-)\Big)\right] + \gamma\right]$$

where $\sigma_{i,j}^{\pm}$ stands for $\sigma^{\pm}(x_j, \theta_{i,j})$ and

$$\gamma = (1 - \alpha_1^*) \cdot \ldots \cdot (1 - \alpha_n^*).$$

Proposition 1: The set of rules $\tilde{b}_i$ such that

$$\alpha_i + \alpha_i^* \leq 1, \qquad i = 1, \dots, n, \tag{14}$$

$$0 \leq \alpha_1^* + \ldots + \alpha_n^* \leq 2 \tag{15}$$

is equal to $(S_0^0 \cup S_1^0) \cup (S_0^1 \cup S_1^1) \cup \ldots \cup (S_0^n \cup S_1^n)$.

Roughly speaking, the first factor of $\tilde{b}$ corresponds to a chain of "and" operations, whereas the second factor includes an "or" operation as a special case (compare with Equations (1)–(2)). If $\alpha_i = \alpha_i^* = 0$, x_i does not appear in $\tilde{b}_i$. If $\alpha_i = 1$, $\sigma^{\pm}(x_i)$ is an "and" factor of $\tilde{b}_i$. Constraint (15) implies that the rightmost factor is either $1 - \sigma^{\pm}(x_j)$ (i.e. $\sigma^{\pm}(x_j)$ is negated and the whole expression is a chain of "and" operations) or $1 - \big(1 - \sigma^{\pm}(x_j)\big)\big(1 - \sigma^{\pm}(x_k)\big)$. The latter case corresponds to $\alpha_j^* = \alpha_k^* = 1$ and encodes an "or" between (possibly negated) x_j and x_k. Finally, $\beta_i = 0$ and $\beta_i = 1$ correspond to having $\sigma_i^-(x_i)$ or $\sigma_i^+(x_i)$, i.e. $\beta_i = 0$ if x_i is negated. The expression of γ makes sure that the second factor of $\tilde{b}_i$ does not vanish when $\alpha_1^* = \ldots = \alpha_n^* = 0$. This parametrization is redundant. For example, an "and" factor $\sigma^+(x_j)$ can be equally rewritten as a "degenerate or" factor $1 - \sigma^-(x_j)$. In the first case, $\alpha_j = 1$ and $\alpha_j^* = 0$; in the latter case, $\alpha_j = 0$ and $\alpha_j^* = 1$. Simulations show that the parametrization is suitable for the optimization problem of interest, see Section VI. We may now restate (13) as the following constrained mixed-integer nonlinear optimization

problem:

$$\min \ \frac{m+p}{m-p} \times \sum_{h=1}^{m} w_h \Big(s_i(t_h) - \tilde{g}_i\big(y_1(t_h), \ldots, y_n(t_h)\big) \Big)^2$$
$$\text{subject to } (14) \text{ and } (15). \tag{16}$$

Minimization is now expressed with respect to α, α^*, β, κ_i^1, κ_i^2 and the sigmoidal parameters $\theta_{i,j}$, with $j = 1, \ldots, n$. The model complexity p is computed as follows:

$$p = 2 + \sum_{j=1}^{n} (1 - (1 - \alpha_j^*)(1 - \alpha_j)) size(\theta_{i,j}),$$

i.e. p counts the number of parameters of the sigmoids included in the model plus the two parameters κ_i^1 and κ_i^2. α, α^* and β are the integer variables of the problem that will be handled with branch-and-bound techniques. For fixed values of α, α^* and β, optimization with respect to the parameters $\theta_{i,j}$ is generally nonconvex. Its complexity depends on the analytic expression of σ^+ and on the number of the unknown parameters p.

We are especially interested in the case of Hill functions. Let $\theta = (\eta, d)$, with $\eta \in \mathbb{R}_+$ and $d \in \mathbb{R}_+$, denote the threshold and cooperativity parameters. Take

$$\sigma^+(x, \theta) = \frac{x^d}{x^d + \eta^d}, \qquad \sigma^-(x, \theta) = \frac{\eta^d}{x^d + \eta^d.} \tag{17}$$

Threshold η is such that $\sigma^+(\eta, \theta) = \sigma^-(\eta, \theta) = 1/2$, whereas d determines how abruptly σ^+ increases from 0 to 1 (σ^- decreases from 1 to 0) as x increases. With this choice, the problem has $3n$ binary variables and $2n + 2$ continuous variables, for a total of $5n + 2$ unknown parameters. A simple modification of the problem statement that reduces the number of binary variables to $2n$ and the total number of parameters to $4n + 2$ is discussed in [2].

VI. NUMERICAL RESULTS

We considered the problem of identifying the model of Example 1 from simulated data. The model is endowed with sigmoidal activation functions in the form (17). Each sigmoid is parameterized by a pair $\theta_i^k = (\eta_i^k, d_i^k)$. The parameter values for this model and the initial conditions were chosen based on comparison with experimental data [3] and can be found in [2]. Note that all equations fall in a class $S_0^\ell \cup S_1^\ell$ with $\ell \leq 6$ except for Eq. (6).

In our experiment we assumed that the cooperativity coefficients are known and fixed to their value ($d = 3$). Since parameters d are not part of the optimization problem, from now on they will be dropped from our notation. In particular, we shall write η in place of θ for the unknown parameters of the sigmoids. We attempt the estimation of the structure of the system along with the synthesis rates and the thresholds of the sigmoidal activation functions. Identification is performed separately for each of the 6 equations of the model and relies on the numerical solution of (16). The total number of unknown parameters for each equation is $4n + 2$. A basic MATLAB implementation of a nonlinear branch-and-bound optimization algorithm, solving appropriate optimization subproblems by continuous relaxation of the integer variables [26], is used for this purpose. Optimization is performed on data from a single simulation of the model. Measurements of protein concentrations and synthesis rates were collected every $T = 5$min over the time interval $[0, 1200]$min, i.e. $t_h = (h-1)T$, with $h = 1, \ldots, m$ and $m = 241$. They were artificially corrupted by zero-mean Gaussian noise generated at random in accordance with the multiplicative noise model. We set $\sigma_\epsilon = 0.01$. This corresponds to a noise magnitude effectively concentrated within 3% of the observed synthesis rate value. For the noise in the observation of x_i we consider two different cases:

1) $\sigma_e = 0$ (i.e. perfect measurements $y_i = x_i$);
2) $\sigma_e = 0.01$ (i.e. noise effectively concentrated within 3% of the observed concentration value).

The comparison of the two cases will reveal the effect of ignoring e in the choice of the regression weights (see Section IV-A). Optimization was initialized by setting each threshold value at the mean value of the corresponding concentration measurements. For each i we initially set $\kappa_i^1 = \min\{s_i(t_h) : \ h = 1, \ldots, m\}$ and $\kappa_i^2 = \max\{s_i(t_h) : \ h = 1, \ldots, m\} - \min\{s_i(t_h) : \ h = 1, \ldots, m\}$. The integer variables α_j and β_j were initially relaxed and set to 0.5 (i.e the mean of the admissible values), whereas the α_j^* were all set to 0 so as to fulfill constraint (15).

Simulation of the model (4)–(9) reveals that some of the sigmoidal nonlinearities are not sufficiently explored by the data. This is due to the fact that the thresholds associated to several sigmoids of the model are not crossed by the protein concentrations they act on. As a result only part of the model can be identified. The identifiable part of the model was defined by a semi-quantitative analysis (see [2]) and turns out to be:

$$g_1 = \kappa_1^1 + \kappa_1^2, \tag{18}$$
$$g_2 = \kappa_2^1 + \kappa_2^2 \sigma^-(x_3, \eta_{2,3}), \tag{19}$$
$$g_3 \simeq \tilde{\kappa}_3^1 + \kappa_3^2 \sigma^+(x_4, \eta_{3,4}) \sigma^-(x_3, \eta_{3,3}), \tag{20}$$
$$g_4 \simeq \kappa_4^2 \sigma^-(x_4, \eta_{4,4}) \sigma^-(x_3, \eta_{4,3}), \tag{21}$$
$$g_5 \simeq \kappa_5^2 \sigma^+(x_4, \eta_{5,4}) \sigma^+(x_3, \eta_{5,3}), \tag{22}$$
$$g_6 = \kappa_6^1 + \kappa_6^2 \sigma^+(x_3, \eta_{6,3}). \tag{23}$$

Note that every function (18)–(23) belongs to $S_0^\ell \cup S_1^\ell$ for some $\ell \leq 6$.

We noticed that model penalization was made less effective by the convergence to local minima corresponding to nonminimal models. To compensate for this, we postprocessed the identification results by the following iterative procedure. Given the identified model, take one sigmoid of the model out and refit the parameters of the remaining sigmoids. Do this iteratively until no sigmoid can be taken out of the model with an improvement in the fitting cost. We verified that this procedure provides simpler models at a lower cost in several cases. This indicates clearly that the nonconvexity of the problem is an issue and that the iterative procedure improves optimization effectively.

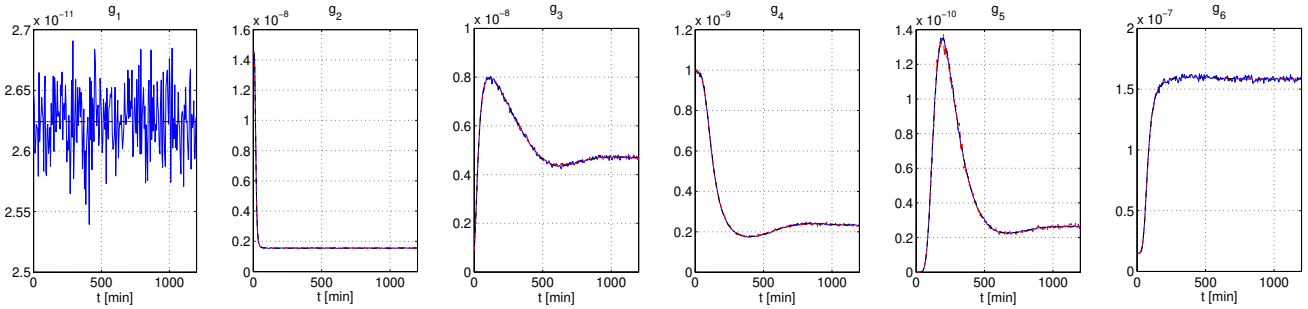

Fig. 1. Optimization results. For $i = 1, \dots 6$, (left to right), each plot reports the matching between the data s_i (blue solid lines), and its estimates $\hat{g}_i(y_1, \dots, y_6)$ with noise on the concentration measurements (red dashed lines). Estimates without noise on concentration measurements overlap with red dashed lines and are not reported.

The interactions identified after postprocessing are

$$\begin{aligned}
\hat{g}_1 =& \hat{\kappa}_1^1 + \hat{\kappa}_1^2 \\
\hat{g}_2 =& \hat{\kappa}_2^1 + \hat{\kappa}_2^2 \sigma^-(x_3, \hat{\eta}_{2,3}) \\
\hat{g}_3 =& \hat{\kappa}_3^1 + \hat{\kappa}_3^2 \sigma^-(x_3, \hat{\eta}_{3,3}) \sigma^+(x_4, \hat{\eta}_{3,4}) \\
\hat{g}_4 =& \hat{\kappa}_4^1 + \hat{\kappa}_4^2 \sigma^-(x_3, \hat{\eta}_{4,3}) \sigma^-(x_4, \hat{\eta}_{4,4}) \\
\hat{g}_5 =& \hat{\kappa}_5^1 + \hat{\kappa}_5^2 \sigma^+(x_3, \hat{\eta}_{5,3}) \sigma^+(x_4, \hat{\eta}_{5,4}) \sigma^-(x_6, \hat{\eta}_{5,6}) \\
\hat{g}_6 =& \hat{\kappa}_6^1 + \hat{\kappa}_6^2 \sigma^+(x_3, \hat{\eta}_{6,3})
\end{aligned}$$

for Case 1 (perfect concentration measurements) and

$$\begin{aligned}
\hat{g}_1 =& \hat{\kappa}_1^1 + \hat{\kappa}_1^2 \\
\hat{g}_2 =& \hat{\kappa}_2^1 + \hat{\kappa}_2^2 \sigma^+(x_2, \hat{\eta}_{2,2}) \sigma^-(x_3, \hat{\eta}_{2,3}) \\
\hat{g}_3 =& \hat{\kappa}_3^1 + \hat{\kappa}_3^2 \sigma^+(x_1, \hat{\eta}_{3,1}) \sigma^-(x_3, \hat{\eta}_{3,3}) \times \\
& \times \sigma^+(x_4, \hat{\eta}_{3,4}) \sigma^-(x_6, \hat{\eta}_{3,6})\big) \\
\hat{g}_4 =& \hat{\kappa}_4^1 + \hat{\kappa}_4^2 \sigma^-(x_4, \hat{\eta}_{4,4}) \sigma^-(x_5, \hat{\eta}_{4,5}) \times \\
& \times \big(1 - \sigma^-(x_2, \hat{\eta}_{4,2}) \sigma^+(x_3, \hat{\eta}_{4,3})\big) \\
\hat{g}_5 =& \hat{\kappa}_5^1 + \hat{\kappa}_5^2 \sigma^+(x_3, \hat{\eta}_{5,3}) \sigma^+(x_4, \hat{\eta}_{5,4}) \\
\hat{g}_6 =& \hat{\kappa}_6^1 + \hat{\kappa}_6^2 \sigma^+(x_3, \hat{\eta}_{6,3})
\end{aligned}$$

for Case 2 (noisy concentration measurements). Superscript " $\hat{}$ " is used to denote estimates. The estimated values of the parameters for the two cases are reported in Tables I and II. Figure 1 shows that the matching between the data and the estimated model predictions is very accurate in all cases.

In general, the estimation of the interactions and of the unknown parameters is quite accurate, but the presence of measurement noise on the concentration values tends to favor the choice of overly complicated models. We interpret this as a consequence of having ignored the contribution of e_i in the choice of the regression weights w_h. In fact, the regression problem could be better formulated as the identification of an errors-in-variables model [10]. The resulting problem is nonstandard due to the nonlinearity of the model and requires further investigation. In Case 1, the estimated structure of the synthesis rates is exact with the exception of $\hat{g}_5$, which contains the spurious sigmoid $\sigma^-(x_6, \hat{\eta}_{5,6})$. Table I shows that the value $\sigma^-(x_6, \hat{\eta}_{5,6})$ is nearly 1 all along the observed state trajectories. An offset in the estimate $\hat{k}_5^2$ can be observed. This is compensated partly by an offset in the estimate $\hat{\eta}_{5,4}$ and partly by the presence of $\sigma^-(x_6, \hat{\eta}_{5,6})$. In both Case 1 and Case 2, there is an identifiability problem for the parameters κ_1^1 and κ_1^2. The reason is clear from the expression of g_1. In both cases, what is estimated accurately is the sum $\hat{\kappa}_1^1 + \hat{\kappa}_1^2 \simeq \kappa_1^1 + \kappa_1^2$. In Case 2, incorrect interactions are introduced in $\hat{g}_3$ and $\hat{g}_4$. For $\hat{g}_3$, Table II suggests that the erroneous presence of $\sigma^-(x_6, \hat{\eta}_{3,6}) \simeq 1$ and of $\sigma^-(x_1, \hat{\eta}_{3,1})$ mitigates the effect of the overestimation of κ_3^2, while the additional error in the estimation of κ_3^1 is masked by the noise and by the remaining terms, which are significantly larger. For $\hat{g}_4$, an interaction with x_5 has been included although $\sigma^-(x_5, \hat{\eta}_{4,5}) \simeq 1$ throughout the experiment. On the other hand, values of $\sigma^-(x_2, \hat{\eta}_{4,2})$ close to but less than 1 appear to balance the underestimation of $\eta_{4,3}$, which causes the values of $\sigma^+(x_3, \hat{\eta}_{4,3})$ to be larger than $\sigma^+(x_3, \eta_{4,3})$. Note that, where $\sigma^-(x_2, \hat{\eta}_{4,2}) \equiv 1$, the contribution of x_2 could be removed and the rightmost factor of $\hat{g}_4$ would reduce to $\sigma^-(x_3, \hat{\eta}_{4,3})$, yielding the true structure of g_4. Since the match between the estimated model predictions and the data is very good, we conclude that the artifacts in the identification results should be attributed in first place to the weak identifiability of certain portions of the model, at least in the given experimental conditions. However, it is clear that more work is also needed to improve the estimation of the model complexity and to explicitly account for the role of the observation noise e_i.

VII. CONCLUSIONS AND PERSPECTIVES

We addressed the identification of gene regulatory networks in a kinetic modelling framework that accounts for the HCF structure of the interactions among genes. Numerical experiments on simulated data showed that our approach is promising, but several issues still need to be addressed. For a fixed model structure, the estimation of the parameters of the sigmoids is a nonconvex problem. At present, the penalization of the model complexity does not provide satisfactory results and makes convergence to local minima more pronounced. Due to the noise in the protein concentration measurements, error-in-variables model identification would be more appropriate and may improve the results. Finally, a theoretical analysis of identifiability and system excitation for kinetic models with HCF structure is needed.

REFERENCES

[1] R. Porreca, S. Drulhe, H. de Jong G. Ferrari-Trecate, "Structural Identification of Piecewise-Linear Models of Genetic Regulatory Net-

Gene	Param	Estimate	True	Sigmoid span
1	κ_1^1	2.477_{10-11}	3.034_{10-12}	
	κ_1^2	1.476_{10-12}	2.317_{10-11}	
2	κ_2^1	1.552_{10-9}	1.553_{10-9}	
	κ_2^2	1.32_{10-8}	1.322_{10-8}	
	$\eta_{2,3}$	4.005_{10-8}	3.991_{10-8}	$[0.000, 0.998]$
3	κ_3^1	3.502_{10-10}	3.404_{10-10}	
	κ_3^2	8.682_{10-9}	8.668_{10-9}	
	$\eta_{3,3}$	1.991_{10-6}	2.020_{10-6}	$[0.807, 1.000]$
	$\eta_{3,4}$	4.02_{10-8}	3.991_{10-8}	$[0.061, 0.910]$
4	κ_4^1	0	0	
	κ_4^2	9.954_{10-10}	9.938_{10-10}	
	$\eta_{4,3}$	7.469_{10-7}	7.472_{10-7}	$[0.181, 1.000]$
	$\eta_{4,4}$	1.883_{10-7}	1.888_{10-7}	$[0.911, 0.999]$
5	κ_5^1	0	0	
	κ_5^2	1.809_{10-9}	2.548_{10-9}	
	$\eta_{5,3}$	7.684_{10-7}	7.472_{10-7}	$[0.000, 0.806]$
	$\eta_{5,4}$	1.63_{10-7}	1.888_{10-7}	$[0.001, 0.131]$
	$\eta_{5,6}$	4.418_{10-5}	–	$[0.929, 1.000]$
6	κ_6^1	1.496_{10-8}	1.506_{10-8}	
	κ_6^2	1.488_{10-7}	1.488_{10-7}	
	$\eta_{6,3}$	3.669_{10-7}	3.663_{10-7}	$[0.000, 0.974]$

TABLE I

ESTIMATION RESULTS FOR CASE 1. For each threshold parameter, the rightmost column indicates the range of values taken on by the estimated sigmoidal nonlinearity along the system trajectory.

Gene	Param	Estimate	True	Sigmoid span
1	κ_1^1	2.477_{10-11}	3.034_{10-12}	
	κ_1^2	1.476_{10-12}	2.317_{10-11}	
2	κ_2^1	1.554_{10-9}	1.553_{10-9}	
	κ_2^2	9.894_{10-7}	1.322_{10-8}	
	$\eta_{2,2}$	2.955_{10-5}	–	$[0.000, 0.014]$
	$\eta_{2,3}$	4.359_{10-8}	3.991_{10-8}	$[0.000, 0.998]$
3	κ_3^1	5.781_{10-10}	3.404_{10-10}	
	κ_3^2	8.797_{10-9}	8.668_{10-9}	
	$\eta_{3,1}$	2.313_{10-10}	–	$[0.529, 0.954]$
	$\eta_{3,3}$	2.116_{10-6}	2.020_{10-6}	$[0.831, 1.000]$
	$\eta_{3,4}$	3.929_{10-8}	3.991_{10-8}	$[0.065, 0.920]$
	$\eta_{3,6}$	4.449_{10-5}	–	$[0.926, 1.000]$
4	κ_4^1	3.967_{10-11}	0	
	κ_4^2	9.634_{10-10}	9.938_{10-10}	
	$\eta_{4,2}$	8.638_{10-6}	–	$[0.644, 1]$
	$\eta_{4,3}$	7.018_{10-7}	7.472_{10-7}	$[0, 0.848]$
	$\eta_{4,4}$	1.613_{10-7}	1.888_{10-7}	$[0.858, 0.999]$
	$\eta_{4,5}$	2.516_{10-8}	(1.221_{10-7})	$[0.879, 1.000]$
5	κ_5^1	0	0	
	κ_5^2	2.252_{10-9}	2.548_{10-9}	
	$\eta_{5,3}$	7.444_{10-7}	7.472_{10-7}	$[0.000, 0.823]$
	$\eta_{5,4}$	1.81_{10-7}	1.888_{10-7}	$[0.001, 0.105]$
6	κ_6^1	1.494_{10-8}	1.506_{10-8}	
	κ_6^2	1.487_{10-7}	1.488_{10-7}	
	$\eta_{6,3}$	3.657_{10-7}	3.663_{10-7}	$[0.000, 0.975]$

TABLE II

ESTIMATION RESULTS FOR CASE 2. For each threshold parameter, the rightmost column indicates the range of values taken on by the estimated sigmoidal nonlinearity along the system trajectory. Threshold values in brackets indicate that the corresponding sigmoid was deemed unidentifiable.

works", *Journal of Computational Biology*, Vol.15, No.10, pp.1365–1380, 2008.

[2] E. Cinquemani, R. Porreca, J. Lygeros and G. Ferrari-Trecate, " Canalizing structure of genetic network dynamics: modelling and identification via mixed-integer programming", Technical report, vol. AUT09-10, ETH Zurich, 2009. `http://control.ee.ethz.ch/index.cgi?page=publications&action=details&id=3358`

[3] H. de Jong and D. Ropers. Personal communication.

[4] D. Ropers, H. de Jong, M. Page, D. Schneider and J. Geiselmann. "Qualitative simulation of the carbon starvation response in *Escherichia coli*", *Biosystems*, Vol.84, No.2, pp.124–152, May 2006.

[5] S. Nikolajewa, M. Friedel and T. Wilhelm. "Boolean networks with biologically relevant rules show ordered behavior", *BioSystems*, 90, pp.40-47, 2007.

[6] S. Kauffman, C. Peterson, B. Samuelsson and C. Troein. "Genetic networks with canalyzing Boolean rules are always stable", *PNAS*, Vol.101, No.49, 17102-17107, December 7, 2004.

[7] Z. Szallasi and S. Liang. "Modeling the normal and neoplastic cell cycle with "realistic boolean genetic networks": their application for understanding carcinogenesis and assessing therapeutic strategies", *Proc. of the 1998 Pacific Symposium on Biocomputing*, 66-76, 1998.

[8] S. Kauffman. "Metabolic stability and epigenesis in randomly constructed genetic nets", *Journal of Theoretical Biology*, 22:437-467, 1969.

[9] H. de Jong, "Modeling and simulation of genetic regulatory systems: A literature review". *Journal of Computational Biology*, 9(1), 69-105, 2002.

[10] T. Söderström and P. Stoica, *System Identification*. Prentice-Hall, 1994.

[11] R. Laubenbacher and B. Stigler, "A computational algebra approach to the reverse engineering of gene regulatory networks", *Journal of Theoretical Biology*, Vol.229, No.4, pp.523-537, 2004.

[12] L. Raeymaekers, *Dynamics of Boolean networks controlled by biologically meaningful functions*, *Journal of Theoretical Biology* Vol.218, No.3, pp.331-342, 2002.

[13] T. Akutsu, S. Miyano, S. Kuhara, "Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function", *J. Comp. Biol.*, Vol.7, No.3-4), p. 331-344.

[14] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller and N. Friedman, "Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data", *Nature Genetics*, Vol.34, No.2, pp.166-176, 2003.

[15] I. Nachman, A. Regev, N. Friedman, "Inferring quantitative models of regulatory networks from expression data", *Bioinformatics*, Vol.20, Suppl. 1, pp.I248-I256, 2004.

[16] H. de Jong, D. Ropers, C. Ranquet, C. Pinel and J. Geiselmann, "Making the Most of Fluorescence and Luminescence Data: The Analysis of High-Precision Measurements of Gene Expression". Submitted.

[17] M.M. Zavlanos, A. Julius, S.P. Boyd and G.J. Pappas, "Identification of stable genetic networks using convex programming", in *Proceedings of the American Control Conference*, Seattle, WA, 2008.

[18] M. Bansal, V. Belcastro, A. Ambesi-Impiombato and D. di Bernardo, "How to infer gene networks from expression profiles", *Molecular Systems Biology*, Vol.3, N.78.

[19] T.S. Gardner, D. di Bernardo, D. Lorenz, J.J. Collins, "Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling", *Science*, Vol.301, N.5629, pp.102-105 2003.

[20] H. de Jong, J.-L. Gouze, C. Hernandez, M. Page, T. Sari and J. Geiselmann, "Hybrid modeling and simulation of genetic regulatory networks: A qualitative approach", in O. Maler and A. Pnueli, Eds., N.2623 of the LNCS Series, pp.267–282, Springer–Verlag, Berlin, 2003.

[21] S. Drulhe, G. Ferrari-Trecate, and H. de Jong. The switching threshold reconstruction problem for piecewise-affine models of genetic regulatory networks. *IEEE Trans. on Circuits and systems I: Regular papers and IEEE Trans. on Automatic Control*, 53:153–165, 2008.

[22] J. Jaeger, S. Surkova, M. Blagov, H. Janssens, D. Kosman, K.N. Kozlov, Manu, E. Myasnikova, C.E. Vanario-Alonso, M. Samsonova, D.H. Sharp and J. Reinitz, "Dynamic control of positional information in the early *Drosophila* embryo", *Nature*, Vol.430, No.6997, pp.368–371, 2004.

[23] E. Cinquemani, A. Milias-Argeitis, S. Summers and J. Lygeros, "Local identification of piecewise deterministic models of genetic networks". In R.Mujumdar and P.Tabuada, Eds., N.5469 of the LNCS Series, pp.105–119, Springer–Verlag, Berlin, 2009.

[24] E. Cinquemani, A. Milias, S. Summers, J. Lygeros, "Stochastic dynamics of genetic networks: modelling and parameter identification". *Bioinformatics*, vol. 24, no. 23, pp. 2748-2754, 2008.

[25] C. Kreutz, M.M. Bartolome Rodriguez, T. Maiwald, M. Seidl, H.E. Blum, L. Mohr, J. Timmer, "An error model for protein quantification". *Bioinformatics*, vol.23, no.20, pp.2747-2753, 2007.

[26] C.A. Floudas, *Nonlinear and Mixed-Integer Optimization*, Orford University Press, New York, 1995.